



A comparison of normalization methods for high density oligonucleotide array data based on variance and bias

B. M. Bolstad^{1,*}, R. A. Irizarry², M. Åstrand³ and T. P. Speed^{4,5}

¹Group in Biostatistics, University of California, Berkeley, CA 94720, USA,

²Department of Biostatistics, John Hopkins University, Baltimore, MD, USA,

³AstraZeneca R & D Mölndal, Sweden, ⁴Department of Statistics, University of California, Berkeley, CA 94720, USA and ⁵Division of Genetics and Bioinformatics, WEHI, Melbourne, Australia

Received on June 13, 2002; revised on September 11, 2002; accepted on September 17, 2002

ABSTRACT

Motivation: When running experiments that involve multiple high density oligonucleotide arrays, it is important to remove sources of variation between arrays of non-biological origin. Normalization is a process for reducing this variation. It is common to see non-linear relations between arrays and the standard normalization provided by Affymetrix does not perform well in these situations.

Results: We present three methods of performing normalization at the probe intensity level. These methods are called complete data methods because they make use of data from all arrays in an experiment to form the normalizing relation. These algorithms are compared to two methods that make use of a baseline array: a one number scaling based algorithm and a method that uses a non-linear normalizing relation by comparing the variability and bias of an expression measure. Two publicly available datasets are used to carry out the comparisons. The simplest and quickest complete data method is found to perform favorably.

Availability: Software implementing all three of the complete data normalization methods is available as part of the R package Affy, which is a part of the Bioconductor project <http://www.bioconductor.org>.

Contact: bolstad@stat.berkeley.edu

Supplementary information: Additional figures may be found at <http://www.stat.berkeley.edu/~bolstad/normalize/index.html>

INTRODUCTION

The high density oligonucleotide microarray technology, as provided by the Affymetrix GeneChip[®], is being used in many areas of biomedical research. As described in Lipshutz *et al.* (1999) and Warrington *et al.* (2000),

oligonucleotides of 25 base pairs in length are used to probe genes. There are two types of probes: reference probes that match a target sequence exactly, called the *perfect match* (PM), and partner probes which differ from the reference probes only by a single base in the center of the sequence. These are called the *mismatch* (MM) probes. Typically 16–20 of these probe pairs, each interrogating a different part of the sequence for a gene, make up what is known as a probeset. Some more recent arrays, such as the HG-U133 arrays, use as few as 11 probes in a probeset. The intensity information from the values of each of the probes in a probeset are combined together to get an expression measure, for example, Average Difference (AvgDiff), the Model Based Expression Index (MBEI) of Li and Wong (2001), the MAS 5.0 Statistical algorithm from Affymetrix (2001), and the Robust Multi-chip Average proposed in Irizarry *et al.* (2003).

The need for normalization arises naturally when dealing with experiments involving multiple arrays. There are two broad characterizations that could be used for the type of variation one might expect to see when comparing arrays: interesting variation and obscuring variation. We would classify biological differences, for example large differences in the expression level of particular genes between a diseased and a normal tissue source, as interesting variation. However, observed expression levels also include variation that is introduced during the process of carrying out the experiment, which could be classified as obscuring variation. Examples of this obscuring variation arise due to differences in sample preparation (for instance labeling differences), production of the arrays and the processing of the arrays (for instance scanner differences). The purpose of normalization is to deal with this obscuring variation. A more complete discussion on the sources of this variation can be found in Hartemink *et al.* (2001).

*To whom correspondence should be addressed.

Affymetrix has approached the normalization problem by proposing that intensities should be scaled so that each array has the same average value. The Affymetrix normalization is performed on expression summary values. This approach does not deal particularly well with cases where there are non-linear relationships between arrays. Approaches using non-linear smooth curves have been proposed in Schadt *et al.* (2001, 2002) and Li and Wong (2001). Another approach is to transform the data so that the distribution of probe intensities is the same across a set of arrays. Sidorov *et al.* (2002) propose parametric and non-parametric methods to achieve this. All these approaches depend on the choice of a baseline array.

We propose three different methods of normalizing probe intensity level oligonucleotide data, none of which is dependent on the choice of a baseline array. Normalization is carried out at probe level for all the probes on an array. Typically we do not treat PM and MM separately, but instead consider them all as intensities that need to be normalized. The normalization methods do not account for saturation. We consider this a separate problem to be dealt with in a different manner.

In this paper, we compare the performance of our three proposed complete data methods. These methods are then compared with two methods making use of a baseline array. The first method, which we shall refer to as the scaling method, mimics the Affymetrix approach. The second method, which we call the non-linear method, mimics the approaches of Schadt *et al.* Our assessment of the normalization procedures is based on empirical results demonstrating ability to reduce variance without increasing bias.

NORMALIZATION ALGORITHMS

Complete data methods

The complete data methods combine information from all arrays to form the normalization relation. The first two methods, cyclic loess and contrast, are extensions of accepted normalization methods that have been used successfully with cDNA microarray data. The third method, based on quantiles, is both quicker and simpler than those methods.

Cyclic loess This approach is based upon the idea of the M versus A plot, where M is the difference in log expression values and A is the average of the log expression values, presented in Dudoit *et al.* (2002). However, rather than being applied to two color channels on the same array, as is done in the cDNA case, it is applied to probe intensities from two arrays at a time. An M versus A plot for normalized data should show a point cloud scattered about the $M = 0$ axis.

For any two arrays i, j with probe intensities x_{ki} and x_{kj} where $k = 1, \dots, p$ represents the probe, we calculate

$M_k = \log_2(x_{ki}/x_{kj})$ and $A_k = \frac{1}{2} \log_2(x_{ki}x_{kj})$. A normalization curve is fitted to this M versus A plot using loess. Loess is a method of local regression (see Cleveland and Devlin 1988 for details). The fits based on the normalization curve are \hat{M}_k and thus the normalization adjustment is $M'_k = M_k - \hat{M}_k$. Adjusted probe intensities are given by $x'_{ki} = 2^{A_k + \frac{M'_k}{2}}$ and $x'_{kj} = 2^{A_k - \frac{M'_k}{2}}$. The preferred method is to compute the normalization curves using rank invariant sets of probes. This paper uses invariants sets since it increases the implementation speed.

To deal with more than two arrays, the method is extended to look at all distinct pairwise combinations. The normalizations are carried out in a pairwise manner as above. We record an adjustment for each of the two arrays in each pair. So after looking at all pairs of arrays for any array k where $1 \leq k \leq n$, we have adjustments for chip k relative to arrays $1, \dots, k-1, k+1, \dots, n$. We weight the adjustments equally and apply to the set of arrays. We have found that after only 1 or 2 complete iterations through all pairwise combinations the changes to be applied become small. However, because this method works in a pairwise manner, it is somewhat time consuming.

Contrast based method The contrast based method is another extension of the M versus A method. Full details can be found in Åstrand (2001). The normalization is carried out by placing the data on a log-scale and transforming the basis. In the transformed basis, a series of $n-1$ normalizing curves are fit in a similar manner to the M versus A approach of the cyclic loess method. The data is then adjusted by using a smooth transformation which adjusts the normalization curve so that it lies along the horizontal. Data in the normalized state is obtained by transforming back to the original basis and exponentiating. The contrast based method is faster than the cyclic method. However, the computation of the loess smoothers is still somewhat time consuming.

Quantile normalization The goal of the quantile method is to make the distribution of probe intensities for each array in a set of arrays the same. The method is motivated by the idea that a quantile-quantile plot shows that the distribution of two data vectors is the same if the plot is a straight diagonal line and not the same if it is other than a diagonal line. This concept is extended to n dimensions so that if all n data vectors have the same distribution, then plotting the quantiles in n dimensions gives a straight line along the line given by the unit vector $(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})$. This suggests we could make a set of data have the same distribution if we project the points of our n dimensional quantile plot onto the diagonal.

Let $\mathbf{q}_k = (q_{k1}, \dots, q_{kn})$ for $k = 1, \dots, p$ be the vector of the k th quantiles for all n arrays $\mathbf{q}_k = (q_{k1}, \dots, q_{kn})$

and $\mathbf{d} = (\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}})$ be the unit diagonal. To transform from the quantiles so that they all lie along the diagonal, consider the projection of \mathbf{q} onto \mathbf{d}

$$\text{proj}_{\mathbf{d}} \mathbf{q}_k = \left(\frac{1}{n} \sum_{j=1}^n q_{kj}, \dots, \frac{1}{n} \sum_{j=1}^n q_{kj} \right)$$

This implies that we can give each array the same distribution by taking the mean quantile and substituting it as the value of the data item in the original dataset. This motivates the following algorithm for normalizing a set of data vectors by giving them the same distribution:

1. given n arrays of length p , form X of dimension $p \times n$ where each array is a column;
2. sort each column of X to give X_{sort} ;
3. take the means across rows of X_{sort} and assign this mean to each element in the row to get X'_{sort} ;
4. get $X_{\text{normalized}}$ by rearranging each column of X'_{sort} to have the same ordering as original X

The quantile normalization method is a specific case of the transformation $x'_i = F^{-1}(G(x_i))$, where we estimate G by the empirical distribution of each array and F using the empirical distribution of the averaged sample quantiles. Extensions of the method could be implemented where F^{-1} and G are more smoothly estimated.

One possible problem with this method is that it forces the values of quantiles to be equal. This would be most problematic in the tails where it is possible that a probe could have the same value across all the arrays. However, in practice, since probeset expression measures are typically computed using the value of multiple probes, we have not found this to be a problem.

Methods using a baseline array

Scaling methods The standard Affymetrix normalization is a scaling method that is carried out on probeset expression measures. To allow consistent comparison with our other methods, we have carried out a similar normalization at the probe level. Our version of this method is to choose a baseline array, in particular, the array having the median of the median intensities. All arrays are then normalized to this 'baseline' via the following method. If x_{base} are the intensities of the baseline array and x_i is any array, then let

$$\beta_i = \frac{\tilde{x}_{\text{base}}}{\tilde{x}_i}$$

where \tilde{x}_i is the trimmed mean intensity (in our analysis we have excluded the highest and lowest 2% of probe

intensities). Then the intensities for the normalized array would be

$$x'_i = \beta_i x_i$$

One can also easily implement the scaling algorithm by using probes from a subset of probesets chosen by using some stability criteria. The HG-U133 arrays provide a set of probesets that have been selected for stability across tissue types, and these could be used for establishing a normalization.

Non-linear method The scaling method is equivalent to fitting a linear relationship with zero intercept between the baseline array and each of the arrays to be normalized. This normalizing relation is then used to map from each array to the baseline array. This idea can be extended to use a non-linear relationship to map between each array and the baseline array. Such an approach is detailed in Schadt *et al.* (2002). This method is used in Li and Wong (2001) and implemented in the dChip software <http://www.dchip.org>. The general approach of these papers is to select a set of approximately rank invariant probes (between the baseline and arrays to be normalized) and fit a non-linear relation, like smoothing splines as in Schadt *et al.* (2002), or a piecewise running median line as in Li and Wong (2001).

The non-linear method used in this paper is as follows. First we select a set of probes for which the ranks are invariant across all the arrays to be normalized. Then we fit loess smoothers to relate the baseline to each of the arrays to be normalized. These loess normalization curves are then used to map probe intensities from the arrays to be normalized to the baseline. This approach is intended to mimic the approach used in dChip. We expect loess smoothers to perform in the same manner as splines or a running median line.

Suppose that $\hat{f}_i(x)$ is the loess smoother mapping from array i to the baseline. Then, in the same notation as above, the normalized array probe intensities are

$$x'_i = \hat{f}_i(x_i)$$

Note that as with the scaling method, the baseline is the array having the median of the median probe intensities.

DATA

We make use of data from two sets of experiments: A dilution/mixture experiment and an experiment using spike-ins. We use these datasets because they allow us to assess bias and variance. The dilution/mixture and spike-in datasets are available directly from GeneLogic (2002) and have been made available for public comparison of analysis methods. This data has been previously described in Irizarry *et al.* (2003).

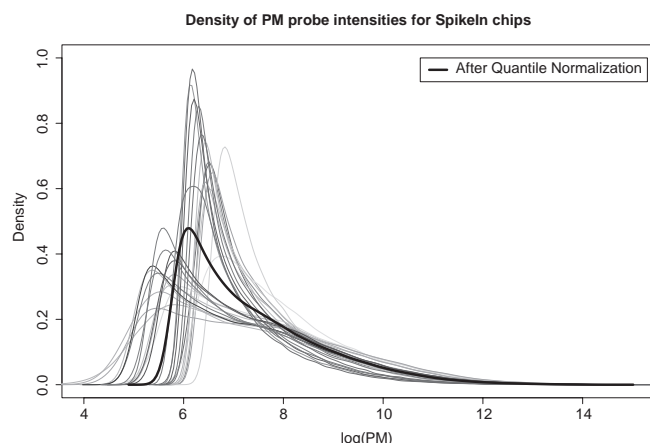


Fig. 1. A plot of the densities for PM for each of the 27 spike-in datasets, with distribution after quantile normalization superimposed.

Dilution/mixture data

The dilution/mixture data series consists of 75 HG-U95A (version 2) arrays, where two sources of RNA, liver (source A), and a central nervous system cell line (source B) are investigated. There are 30 arrays for each source, broken into 6 groups at 5 dilution levels. The remaining 15 arrays, broken into 3 groups of 5 chips, involve mixtures of the two tissue lines in the following proportions: 75 : 25, 50 : 50, and 25 : 75.

Spike-in data

The spike-in data series consists of 98 HG-U95A (version 1) arrays where 11 different cRNA fragments have been spiked in at various concentrations. There is a dilution series consisting of 27 arrays which we will examine in this paper. The remaining arrays are two sets of latin square experiments, where in most cases three replicate arrays have been used for each combination of spike-in concentrations. We make use of 6 arrays (two sets of triplicates) from one of the latin squares.

RESULTS

Probe level analysis

Figure 1 plots the densities for the $\log(PM)$ for each of the 27 arrays from the spike-in dataset, along with the distribution obtained after quantile normalization.

An M versus A plot allows us to discern intensity dependent differences between two arrays. Figure 2 shows M versus A plots for unadjusted PM for all 10 possible pairs of 5 arrays in the liver 10 group before normalization. Clear differences between the arrays can be seen by looking at the loess lines. The point clouds are not centered around $M = 0$ and we see non-linear relationships

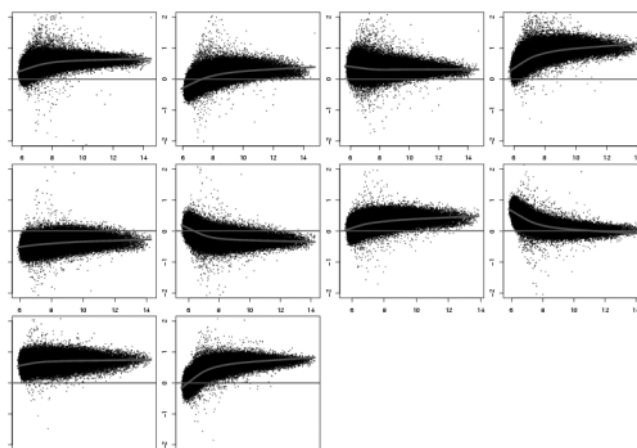


Fig. 2. 10 pairwise M versus A plots using liver (at concentration 10) dilution series data for unadjusted data.

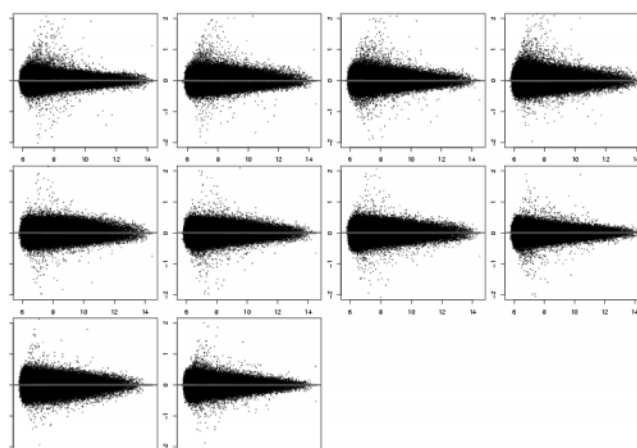


Fig. 3. 10 pairwise M versus A plots using liver (at concentration 10) dilution series data after quantile normalization.

between arrays. The same 10 pairwise comparisons can be seen after quantile normalization in Figure 3. The point clouds are all centered around $M = 0$. Plots produced using the contrast and cyclic loess normalizations are similar.

Expression measures

Comparing normalization methods at the probeset level requires that one must decide on an expression measure. Although in this paper we focus only on one expression measure, the results obtained are similar when using other measures.

The expression summary used in this paper is a robust combination of background adjusted PM intensities and is outlined in Irizarry *et al.* (2003). We call this method the Robust Multichip Average (RMA). RMA estimates

are based upon a robust average of $\log_2(B(PM))$, where $B(PM)$ are background corrected PM intensities. The expression measure may be used on either the natural or log scales.

Irizarry *et al.* (2003) contains a more complete discussion of the RMA measure, and further papers exploring its properties are under preparation.

Probeset measure comparisons

Variance comparisons In the context of the dilution study, consider the five arrays from a single RNA source within a particular dilution level. We calculate expression measures for every probeset on each array and then compute the variance and mean of the probeset expression summary across the five arrays. This is repeated for each group of 5 arrays for the entire dilution/mixture study. We do this after normalization by each of our three complete data methods.

Plotting the log of the ratio of variances versus the average of the log of the mean (expression measure across arrays) allows us to see differences in the between array variations and intensity dependent trends when comparing normalization methods. In this case, the expression measures have all been calculated on the natural scale. Figure 4 shows such plots for the liver at the dilution level 10. Specifically, the four plots compare the variance ratios for quantile : unnormalized, loess : quantile, contrast : quantile and contrast : loess. The horizontal line indicates the x -axis. The other line is a loess smoother. Where the loess smoother is below the x -axis, the first method in the ratio has the smaller ratio and vice versa when the loess smoother is above the line. All three methods reduce the variance at all intensity levels in comparison to data that has not been normalized. The three normalization methods perform in a relatively comparable manner, but the quantile method performs slightly better for this dataset, as can be seen in the loess : quantile and contrast : quantile plots. Similar results are seen in comparable plots (not shown) for the other dilution/mixture groups.

We repeat this analysis with the 27 spike-in arrays, but this time we include the two baseline methods in our comparison. The complete data methods generally leave the mean level of a particular probeset at a level similar to that achieved when using unnormalized data. However, when one of the two baseline methods is used, the mean of a particular probeset is more reminiscent of the value of that probeset in the baseline array. In the natural scale, it is easy to see a mean-variance relationship, where a higher mean implies high variability. Thus, when a comparison is made between the baseline methods and the complete data methods, we find that if a baseline array which shifts the intensities higher (or lower) than the level of those of the unnormalized means is selected, then the corresponding variance of the probeset measures across arrays is higher

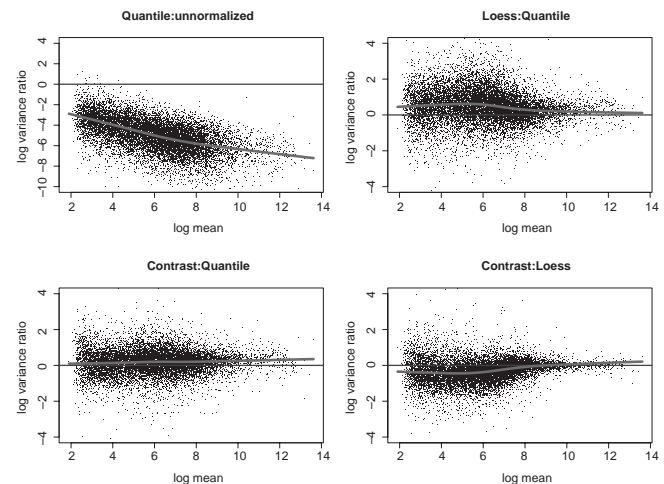


Fig. 4. \log_2 variance ratio versus average \log_2 mean for liver dilution data at concentration 10.

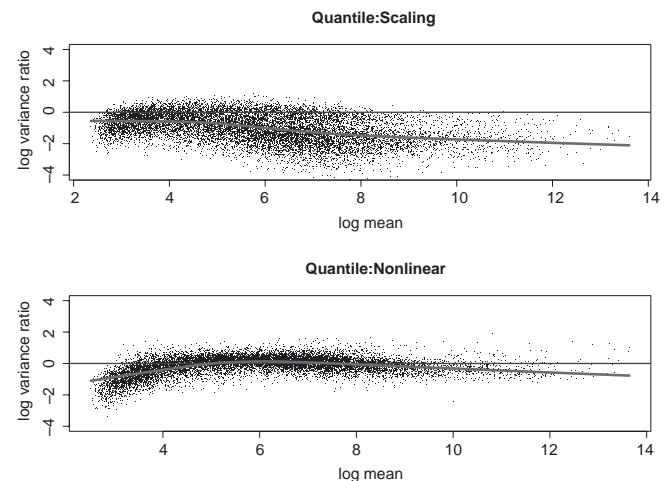


Fig. 5. \log_2 variance ratio versus average \log_2 mean using the spike-in data. Comparing the baseline methods with the quantile method.

(or lower) due to the shifting and not because of the normalization. To minimize this problem and make a fairer comparison, we work with the expression measure on the log scale when comparing the baseline methods to the complete data methods. Figure 5 compares the baseline methods to the quantile methods. We see that the quantile method reduces the between array variances more than the scaling method. The non-linear normalization performs a great deal closer to the quantile method. Similar plots (not shown) comparing the complete data methods with each other for the spike-in data demonstrate that quantile normalization has a slight edge over all the other methods.

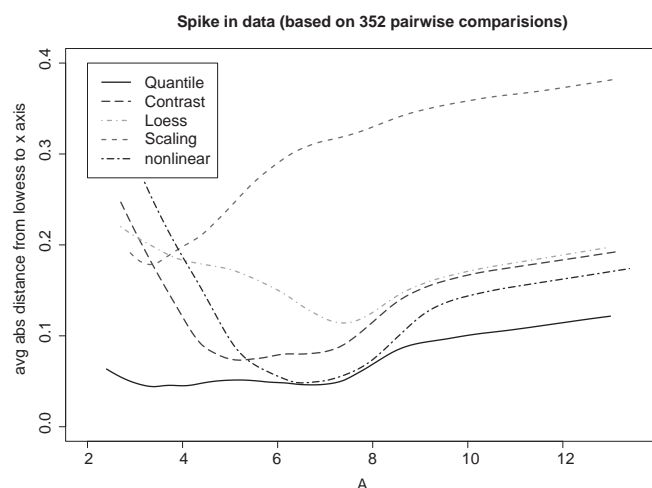


Fig. 6. Comparing the ability of methods to reduce pairwise differences between arrays by using average absolute distance from loess smoother to x -axis in pairwise M versus A plots using spike-in dataset. Smaller distances are favorable.

A similar plot (not shown) comparing the two baseline methods shows as expected that the non-linear method reduces variance when compared to the scaling method.

Pairwise comparison The ability to minimize differences in pairwise comparisons between arrays is a desirable feature of a normalization procedure. An M versus A plot comparing expression measures on two arrays should be centered around $M = 0$ if there is no clear trend towards one of the arrays. Looking at the absolute distance between a loess for the M versus A plot and the x -axis allows us to assess the difference in array to array comparisons. We can compare methods by looking at this distance across a range of intensities and averaging the distance across all pairwise comparisons.

Figure 6 shows such a plot for the spike-in data, we see that the scaling method performs quite poorly when compared to the three complete data methods. The non-linear method performs at a similar level to the complete data methods. For this dataset the quantile method is slightly better. An important property of the quantile method is that these differences remain relatively constant across intensities.

Bias comparisons One way to look at bias is in the context of the spike-in dilution series. We use data for the 27 arrays from the spike-in experiment with 11 control fragments spiked in at 13 different concentrations (0.00, 0.50, 0.75, 1.00, 1.50, 2.00, 3.00, 5.00, 12.50, 25.00, 50.00, 75.00, 100.00, 150.00 pM). We normalize each of the 27 arrays as a group using each of the quantile, contrast, cyclic loess and scaling normalizations. To the

spike-in probesets, we fit the following linear model

$$\log_2 E = \beta_0 + \beta_1 \log_2 c + \epsilon$$

where E is the value of the expression measure and c are the concentrations. Note that the array with spike-in concentration 0 is excluded from the model fit, although it is used in the normalization. The ideal results would be to have slopes that are near 1. Table 1 shows the slope estimates for each of the spike-in probesets after normalization by each of the three complete data methods, the two methods using a baseline and when no normalization has taken place. For the three complete data methods for 10 out of the 11 spike-in probesets, the quantile method gives a slope closer to 1 and the non-linear method has slopes lower than the complete data methods. However, both the scaling and not normalizing have slopes closer to 1. For the non-spike-in probesets on these arrays we should see no linear relation if we fit the same model, since there should be no relation between the spike-in concentrations and the probeset measures. Fitting the linear model above, we find that there is a median slope of 0.042 for these probesets using the unnormalized data. For the quantile method the value of the median slope is -0.005 . All the other normalization methods have median slope near 0. This is about the same difference in slopes as we observed for the spike-ins when comparing the unnormalized data and the best of the normalization methods. In other words, there is a systematic trend due to the manner in which the arrays were produced that has resulted in the intensities of all the probesets being related to the concentration of the spike-ins. We should adjust the spike-in slopes by these amounts. For example, we could adjust the slope of BioB-5 for the quantile method to $0.845 + 0.005 = 0.850$ and the unnormalized slope to $0.893 - 0.042 = 0.851$.

The average R^2 for the spike-in probesets, excluding CreX-3, are 0.87 for the quantile method, and 0.855, 0.849, 0.857 and 0.859 for the contrast, loess, non-linear and scaling methods, respectively. It was 0.831 for the unnormalized data. The median standard error for the slopes was 0.063 for the quantile method. For the other methods these standard errors were 0.065 (contrast), 0.068 (cyclic loess), 0.063 (non-linear), 0.065 (scaling) and 0.076 (unnormalized). Thus of all the algorithms, the quantile method has high slopes, a better fitting model and more precise slope estimates.

The slopes may not reach 1 for several reasons. It is possible that there is a 'pipette' effect. In other words we can not be completely sure of the concentrations. It is more likely that we observe concentration plus an error which leads to a downward bias in the slope estimates. Other possible reasons include the saturation of signal at the high end (this is not a concern with this data) and having a higher background effect at the lower end.

Table 1. Regression slope estimates for spike-in probesets. A slope closer to one is better

Name	Quantile	Contrast	Loess	Non-linear	Scaling	None
AFFX-BioB-5_at	0.845	0.837	0.834	0.803	0.850	0.893
AFFX-DapX-M_at	0.778	0.771	0.770	0.746	0.783	0.826
AFFX-DapX-5_at	0.754	0.747	0.728	0.731	0.764	0.807
AFFX-CreX-5_at	0.903	0.897	0.889	0.875	0.912	0.955
AFFX-BioB-3_at	0.836	0.834	0.825	0.807	0.848	0.890
AFFX-BioB-M_at	0.789	0.782	0.781	0.762	0.797	0.838
AFFX-BioDn-3_at	0.547	0.543	0.550	0.514	0.553	0.595
AFFX-BioC-5_at	0.801	0.794	0.793	0.763	0.808	0.851
AFFX-BioC-3_at	0.796	0.790	0.785	0.769	0.805	0.847
AFFX-DapX-3_at	0.812	0.804	0.793	0.776	0.815	0.859
AFFX-CreX-3_at	−0.007	−0.006	0.002	−0.007	0.005	0.046
Non-spike-in (median)	−0.005	−0.005	−0.005	−0.007	−0.001	0.042

The problems with choosing a baseline The non-linear (and scaling) method requires the choice of a baseline. In this paper we have chosen the array having the median median, but other options are certainly possible. To address these concerns we examine a set of six arrays chosen from the spike-in datasets. In particular, we choose two sets of triplicates, where the fold change of each of the spike-in probesets between the two triplicates is large. The two triplicates are chosen so that about half of the probesets are high in one triplicate and low in the other and vice versa.

We normalize this dataset using both the quantile and non-linear methods. However, for the non-linear method we experiment with the use of each of the six arrays as the baseline array. We also try using two synthetic baseline arrays: one constructed by taking probewise means and one taking probewise medians. Figure 7 shows the distribution of the mean of the probeset measure across arrays. We see that the quantile normalization produces a set of means that is very similar in distribution to the means of the unnormalized data. The means from the non-linear normalizations using each of the six different baseline arrays are quite different from each other and the unnormalized data. This is somewhat of a drawback to the baseline methods. It seems more representative of the complete data to consider all arrays in the normalization rather than to use only a single baseline and give the normalized data characteristics closer to those of one particular array. Only the mean based synthetic baseline array comes close to the unnormalized and quantile methods.

Table 2 summarizes some results from this analysis. We see that all the methods reduce the variability of the probeset measure between arrays compared to that of the unnormalized data. In each case, around 95% of the probesets have reduced variance. When compared to the quantile method, it comes out more even with a little over 50% of probesets having reduced variance for four of

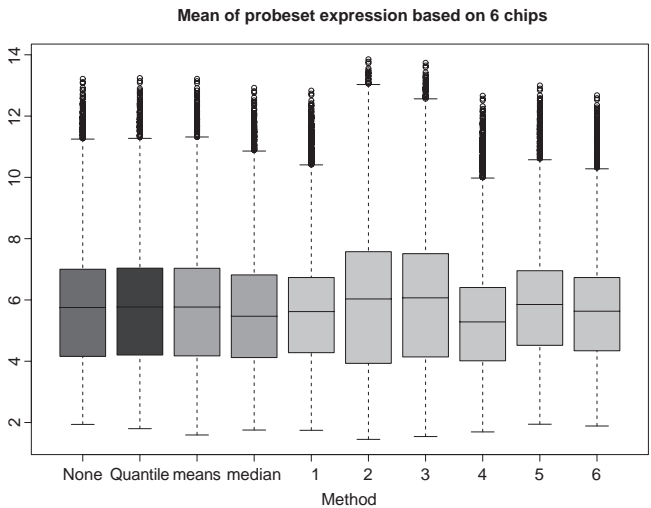


Fig. 7. Distribution of average (over 6 chips) of a probeset expression measure using different baseline normalizations.

the baselines. However, two baseline arrays perform quite poorly. As noted before, this is a reflection of the baseline methods shifting the intensities higher or lower depending on the baseline. The mean based synthetic baseline does not reduce the variability of the probeset measure to the same degree as the quantile method.

Looking at the 11 spike-in probesets, we calculate bias by taking the difference between the log of the ratio between spike-in concentrations and the log of the ratio of intensities in the two groups. One spike-in, Crex-3, did not seem to perform quite as well as the other spike-ins and was excluded from the analysis. Looking at the total absolute bias across the 10 spike-in probesets, we see that the non-linear method has lower total bias (compared to the quantile method) for four of the methods, but two

Table 2. Comparing variance and bias with the non-linear normalization when using different baselines

Method	% with lower var reduced cf. U	% lower var reduced cf. Q	Abs Bias	# abs Bias cf U	# abs Bias cf Q
Probewise mean	83	40	9.2	5	5
Probewise median	96	58	7.9	6	6
Non-linear 1	96	53	7.5	7	5
Non-linear 2	93	31	11.8	2	4
Non-linear 3	94	37	10.5	4	4
Non-linear 4	95	47	7.4	6	5
Non-linear 5	96	55	7.4	7	5
Non-linear 6	96	55	7.5	7	5
Quantile (Q)	95	NA	8.5	6	NA
Unnormalized (U)	NA	NA	9.7	NA	NA

are even bigger than for unnormalized data. Again, this is related to the mean–variance relationship. The four baselines shifted slightly lower in the intensity scale give the most precise estimates. Using this logic, one could argue that choosing the array with the smallest spread and centered at the lowest level would be the best, but this does not seem to treat the data on all arrays fairly. Compared to the unnormalized data, 6 of the spike-in probesets from the quantile normalized data have a smaller bias. For the non-linear normalization using array 1 as the baseline (this is the array chosen using our heuristic), 7 had smaller bias. However, looking at the other baselines, anywhere from 2 to 7 probesets had lower bias. When compared to the quantile method, the results are more even, with about an equal number of the spike-in probesets having a lower bias when using the non-linear method as when using the quantile normalization. An *M* versus *A* plot between the two groups shows all the spike-in points clearly outside the point cloud, no matter which normalization is used. This plot for quantile normalized data is shown in Figure 8.

CONCLUSIONS

We have presented three complete data methods of normalization and compared these to two different methods that make use of a baseline array. Using two different datasets, we established that all three of the complete data methods reduced the variation of a probeset measure across a set of arrays to a greater degree than the scaling method and unnormalized data. The non-linear method seemed to perform at a level similar to the complete data methods. Our three complete data methods, while different, performed comparably at reducing variability across arrays.

When making pairwise comparisons the quantile method gave the smallest distance between arrays. These distances also remained fairly constant across intensities.

In relation to bias, all three complete data methods

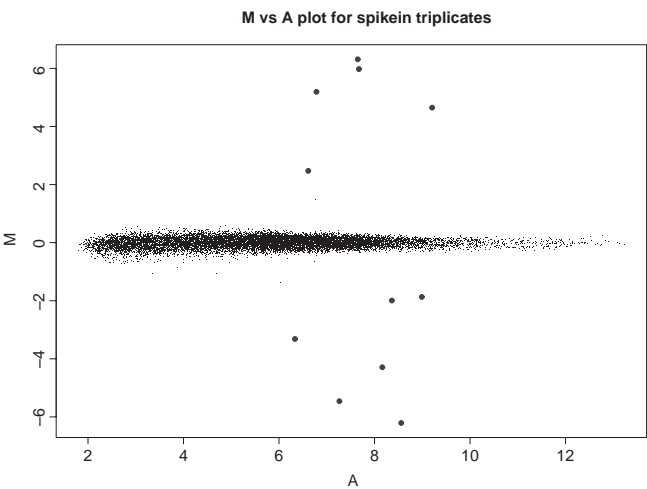


Fig. 8. *M* versus *A* plot for spike-in triplicate data normalized using quantile normalization. Spike-ins are clearly identified.

performed comparably, with perhaps a slight advantage to the quantile normalization. The non-linear method did poorer for the spike-in regressions. The scaling method had slightly higher slopes. Even so, they were more variable.

We saw that the choice of a baseline does have ramifications on down-stream analysis. Choosing a poor baseline would conceivably give poorer results. We also saw that the complete data methods perform well at both variance reduction and on the matter of bias, and in addition more fully reflect the complete set of data. For this reason we favor a complete data method.

In terms of speed, for the three complete data methods, the quantile method is the fastest. The contrast method is slower and the cyclic loess method is the most time consuming. The contrast and cyclic loess algorithms are modifications of an accepted method of normalization.

The quantile method has performed favorably, both in terms of speed and when using our variance and bias criteria, and therefore should be used in preference to the other methods.

While there might be some advantages to using a common, non-data driven, distribution with the quantile method, it seems unlikely an agreed standard could be reached. Different choices of a standard distribution might be reflected in different estimated fold changes. For this reason we prefer the minimalist approach of a data based normalization.

REFERENCES

- Affymetrix (2001) Statistical algorithms reference guide, Technical report, Affymetrix.
- Åstrand, M. (2001) Normalizing oligonucleotide arrays. *Unpublished Manuscript*. <http://www.math.chalmers.se/~magnusaa/maffy.pdf>
- Bolstad, B. (2001) Probe level quantile normalization of high density oligonucleotide array data. *Unpublished Manuscript*. <http://www.stat.berkeley.edu/~bolstad/>
- Cleveland, W.S. and Devlin, S.J. (1998) Locally-weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.*, **83**, 596–610.
- Dudoit, S., Yang, Y.H., Callow, M.J. and Speed, T.P. (2002) Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. *Stat. Sin.*, **12**(1), 111–139.
- GeneLogic (2002) Datasets <http://www.genelogic.com>.
- Hartemink, A., Gifford, D., Jaakkola, T. and Young, R. (2001) Maximum likelihood estimation of optimal scaling factors for expression array normalization. In *SPIE BIOS 2001*.
- Ihaka, R. and Gentleman, R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**(3), 299–314.
- Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K., Scherf, U. and Speed, T.P. (2003) Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics*. in press.
- Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error applications. *Genome Biol.*, **2**(8), 1–11.
- Lipshutz, R., Fodor, S., Gingeras, T. and Lockart, D. (1999) High density synthetic oligonucleotide arrays. *Nature Genet.*, **21**(Suppl), 20–24.
- Schadt, E., Li, C., Su, C. and Wong, W.H. (2001) Analyzing high-density oligonucleotide gene expression array data. *J. Cell. Biochem.*, **80**, 192–202.
- Schadt, E., Li, C., Eliss, B. and Wong, W.H. (2002) Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *J. Cell. Biochem.*, **84**(S37), 120–125.
- Sidorov, I.A., Hosack, D.A., Gee, D., Yang, J., Cam, M.C., Lempicki, R.A. and Dimitrov, D.S. (2002) Oligonucleotide microarray data distribution and normalization. *Information Sciences*, **146**, 65–71.
- Venables, W. and Ripley, B.D. (1997) *Modern Applied Statistics with S-PLUS*, Second edn, Springer, New York.
- Warrington, J.A., Dee, S. and Trulson, M. (2000) Large-scale genomic analysis using Affymetrix GeneChip®. In Schena, M. (ed.), *Microarray Biochip Technology*. BioTechniques Books, New York, **Chapter 6**, pp. 119–148.